Differentially Regulated Pathway Analysis of RNA-Seq by Deep Learning

Huijing Wang[1] and Christian Ray[2]

***Short Abstract* — High throughput RNA sequencing technology is widely used as a tool for transcriptomic analysis in recent years. Large network inference from traditional RNA-Seq differential expression analysis can be limited and biased by researcher subjectivity. As deep learning plays an advantageous role in motif recognition and classification for big data, here we propose training artificial neural networks (ANN) on RNA-Seq abundance data for studying a bacterial phenotypic transition into the persister state. Automated pathway activation recognition by ANN may potentially provide an efficient, accurate, and adaptive tool for network inference.**

***Keywords* — High throughput, RNA-Seq, data inference, deep learning, neural network**

I. Introduction

Bacteria are everywhere and integral to human life. Slow growing pathogenic bacteria are able to tolerate long-term treatment without a cure because of so-called persister cells. Contrary to the common mechanism of stringent response-mediated persister formation, our lab's previous findings show that a novel type of persister occurs in the *E. coli* strain B REL606 when cultured in minimal media containing excess lactose. This effect may arise from fluctuations in critical metabolic pathways [1].

High throughput RNA sequencing technology is essential to biological fields such as epigenomics and transcriptomics [2]. As machine learning (ML) has rapidly evolved into a popular tool in many fields, ML implementations for analyzing RNA-Seq are being developed for sequence analysis and differential expression analysis [3, 4]. However, no implementation of ML is available for pathway analysis, which is necessary for identifying regulatory and metabolic pathways involved in novel phenotypes such as lactose-mediated persister formation. We applied deep learning to unveil the underlying pathway dispersion responsible for the novel persister phenotype.

II. Methods

Unlike image data, transcription counts are independent of data orientation, but connected by the underlying regulatory network. The counts are affected by multiple factors, such as noise, enzyme efficiency, and gene regulation. Here we use TensorFlow [5], a python ML package, to create ANN for RNA-Seq count profile.

A. ANN for identifying differentially regulated pathways (DRP)

We used two methods for DRP detection.

The first method involves using KEGG, a knowledge-based pathway database [6]. We trained an ANN to identify known KEGG pathways and modules. This classifier probabilistically recognizes or rejects gene clusters that have similar transcriptional patterns without explicitly considering regulatory network topology because the pathways and modules are given.

The second method considers regulatory network topology as a component for depicting pathway interactions. Without knowledge-based pathway information, we train the ANN to identify different network motifs with increasing complexity. In this way, comparing existing pathways with the machine-learned pathways, we can validate the ANN as well as potentially discover unknown pathway interactions.

B. Non-ML differential expression analysis

We also performed non-ML DRP analysis by locating differentially expressing genes, which might be responsible for bacterial phenotypic switches in known regulatory pathways in KEGG. This step allows us to compare our machine learning methods with well-established RNA-Seq analysis methods for large network inference.

III. Conclusion

Deep learning by artificial neural network is an unbiased method for processing high-dimensional reduction problems. Once a proper ANN is established for cell expression profiles, it can reveal hidden regulatory patterns in large networks, with which we expect to find clues about bacterial persister formation.

References

[1] J.C.J. Ray *et al*. (2016) Cellular Growth Arrest and Persistence from Enzyme Saturation. PLoS Comput Biol 12(3):e1004825.

[2] J.A. Reuter, D. Spacek, M.P. Snyder. (2015) High-Throughput Sequencing Technologies. Mol Cell 58(4):586-597.

[3] A. Jabeen, N. Ahmad, K. Raza. (2018) Machine Learning-Based State-Of-The-Art Methods For The Classification Of RNA-Seq Data. Classification in BioApps. Springer International Publishing p. 133-172.

[4] I. Kuznetsova *et al*. (2017) Review of Machine Learning Algorithms in Differential Expression Analysis. International SERIES on Information Systems and Management in Creative eMedia (CreMedia). 2016/2, p. 11-24. ISSN 2341-5576

[5] M. Abadi *et al*. (2016) TensorFlow: A system for large-scale machine learning. OSDI 16.

[6] M. Kanehisa, S. Goto. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Research 28(1) 27-30.

Center for Computational Biology, University of Kansas.
[1]E-mail: h749w664@ku.edu. [2]E-mail: jjray@ku.edu.
This work was funded by NIH grant P20GM103638.